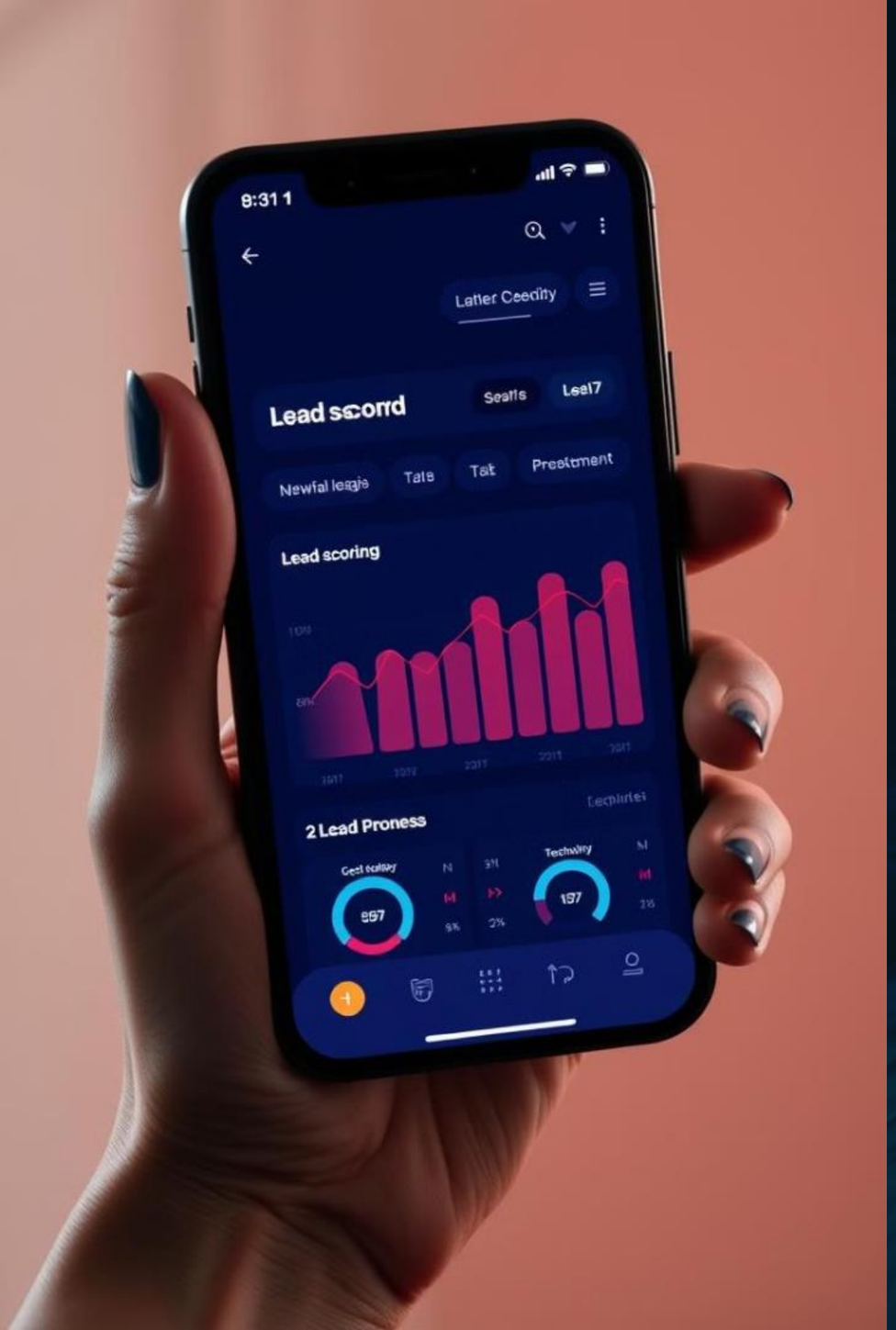


Lead Scoring Case Study

This case study explores the process of building a lead scoring model to model to identify high-potential leads for a B2B company.

A **by Arup Ghosh**

Problem Statement

1. X Education sells online courses to industry Professionals.

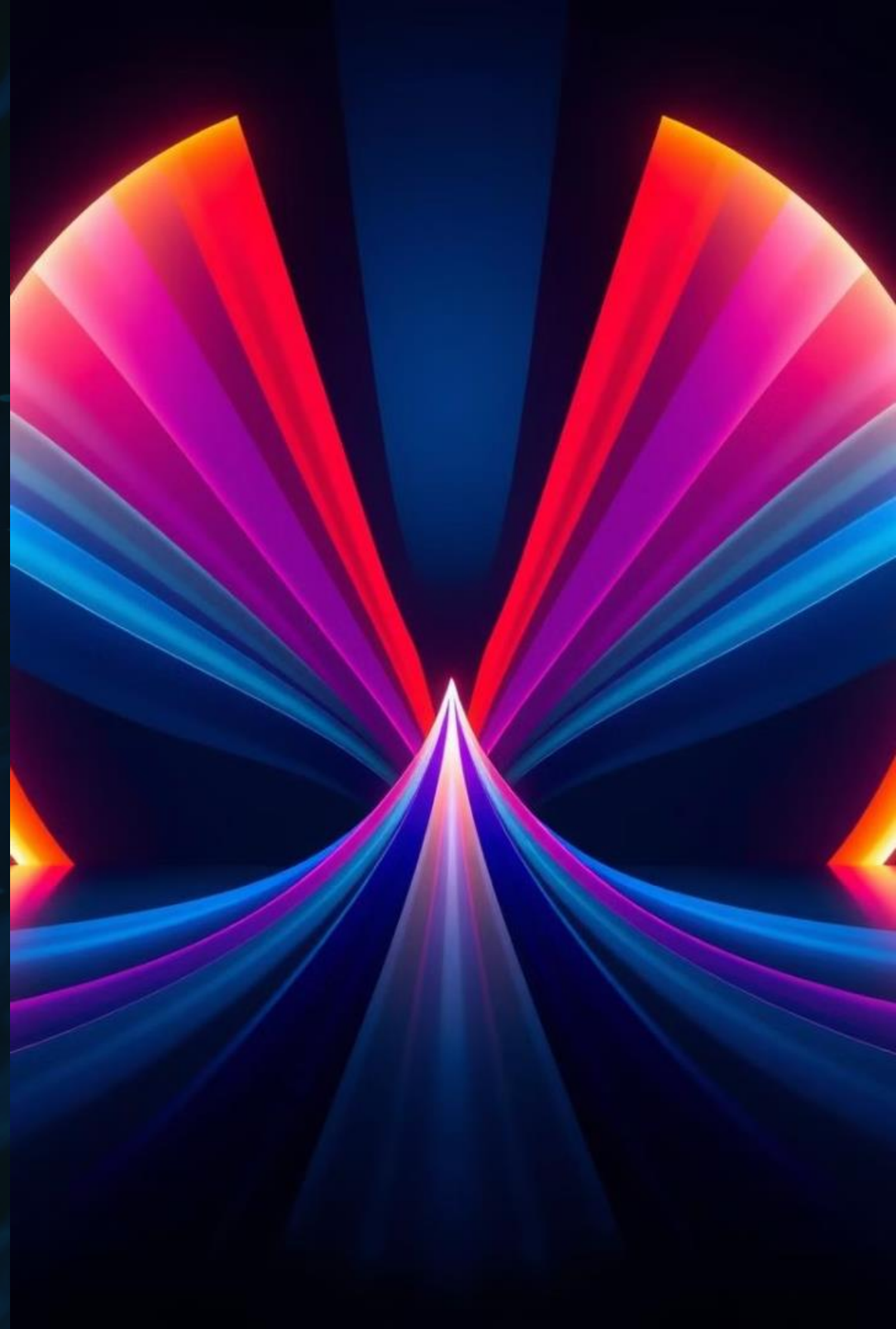
2. X Education gets a lot of leads, its lead conversion rate is Very poor. For example, if, say, they acquire 100 leads in a day, only about 30 of them are converted.

3. To make this process more efficient, the company wishes To identify the most potential leads, also known as ‘Hot Leads’.

4. If they successfully identify this set of leads, the lead conversion rate should go upas the sales team will now be focusing more on communicating with the potential leads rather than making calls to everyone.

Business Objective

- X Education wants to know most promising leads.
- For that they want to build a Model which identifies the hot leads.
- Deployment of the model for the future use.



Lead Scoring Solution Methodology

The lead scoring methodology involved assigning points to leads based on their characteristics and behaviors.

1

Data Cleaning and data manipulation

Check and handle the duplicate data, handle NA and missing values, drop columns if it contains a large number of missing values, imputation of the values, check and handle the outliers in data

2

EDA & Feature Scaling

Transforming raw data into meaningful features that can be used to predict lead quality. (Univariate & Bivariate data analysis, feature scaling and dummy variables and encoding the data.)

3

Model Training

Training a machine learning model to learn the relationship between features and lead conversion.

4

Model Evaluation

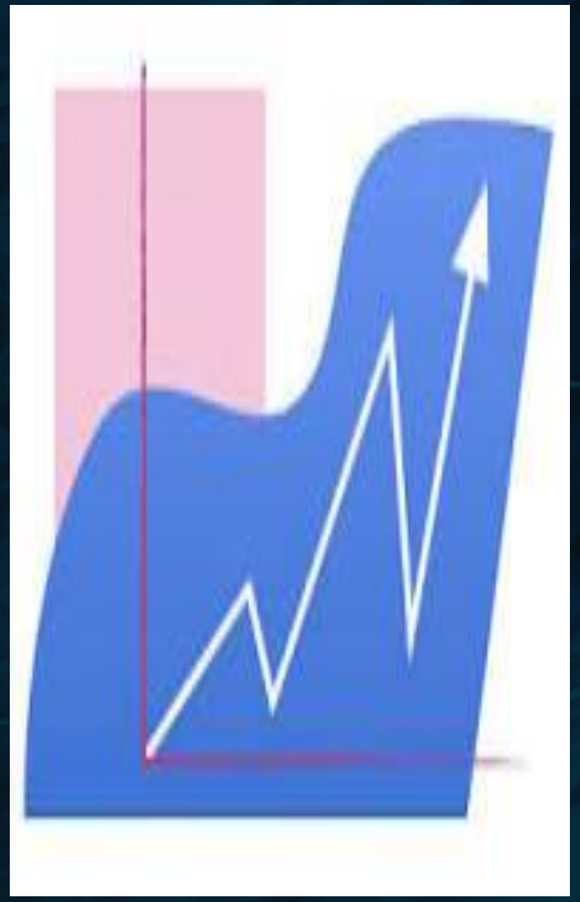
Evaluating the model's performance using metrics such as accuracy, precision, and recall.

5

Deployment & Monitoring

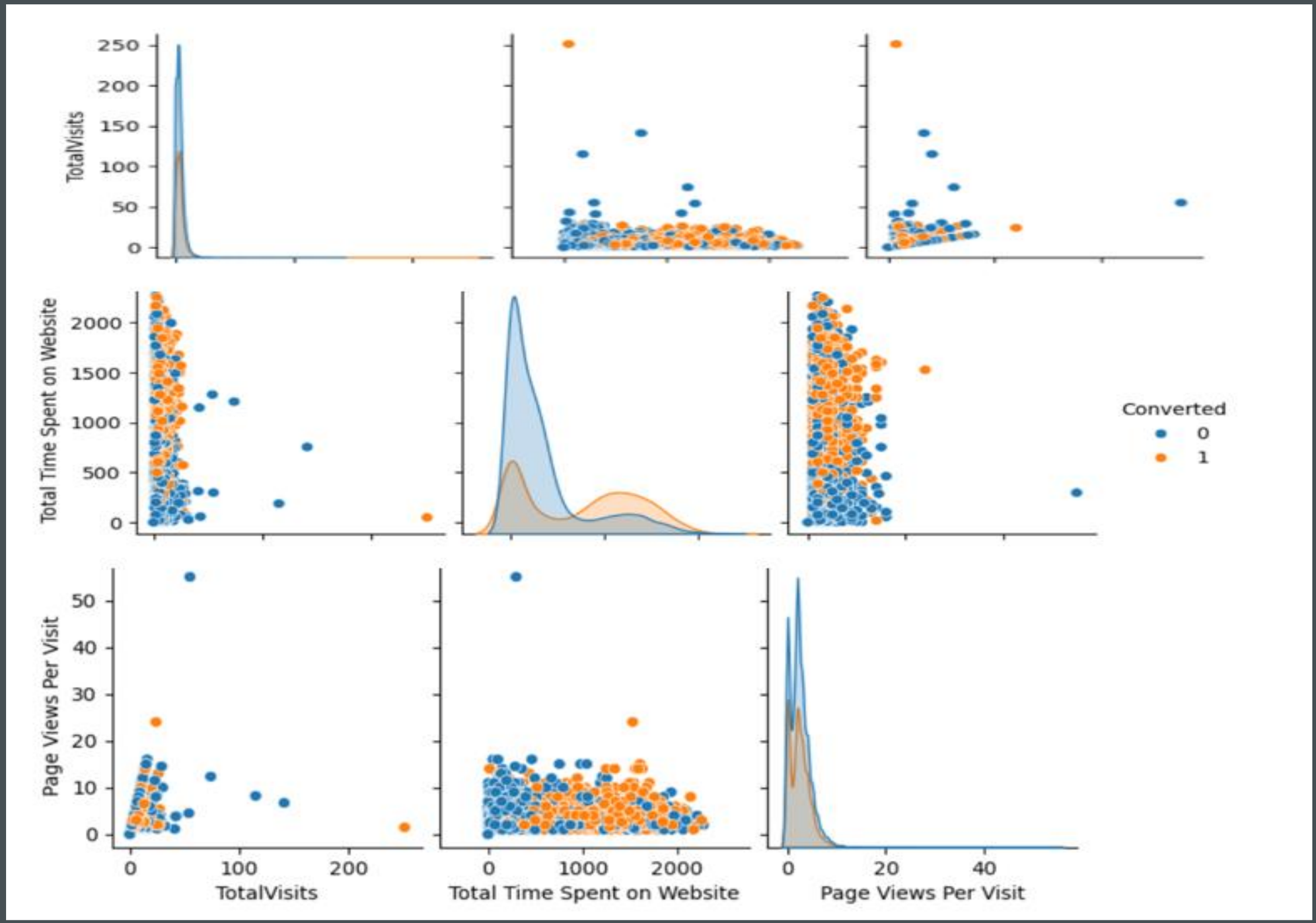
Deploying the model into production and continuously monitoring its performance.

DATA MANIPULATION

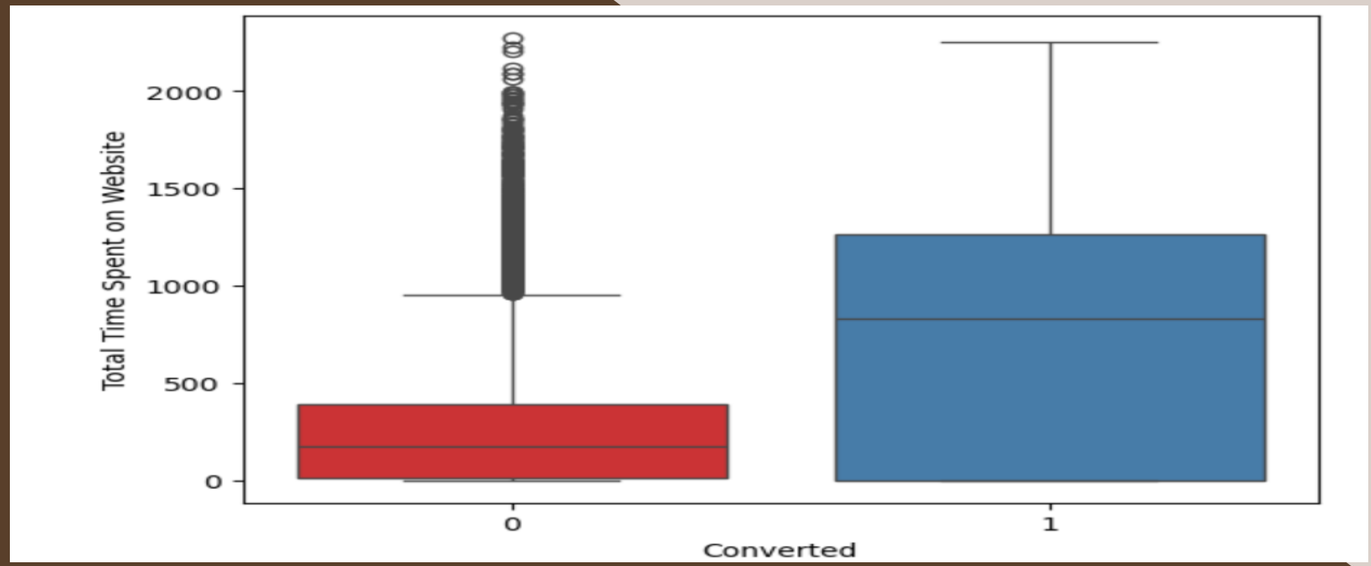


- Total Number of Rows=37,Total Number of Columns =9240.
- Single value features like “Magazine”, “Receive More Updates About Our Courses”, “Update my supply”
- Chain Content”, “Get updates on DM Content”, “I agree to pay the amount through cheque” etc. have been dropped
- Removing the “Prospect ID” and “Lead Number” which are not necessary for the analysis.
- After checking for the value counts for some of the object type variables, we find some of the features which have enough variance, which have dropped, the features are: “Do Not Call”, “What matters most to you in choosing course”, “Search”, “Newspaper, Article”, “X Education Forums”, “Newspaper”, “Digital Advertisement” etc.
- Dropping the column shaving more than 35% as missing values such as ‘How did you Hear about X Education’ and ‘Lead Profile’.

Exploratory Data Analysis (EDA)



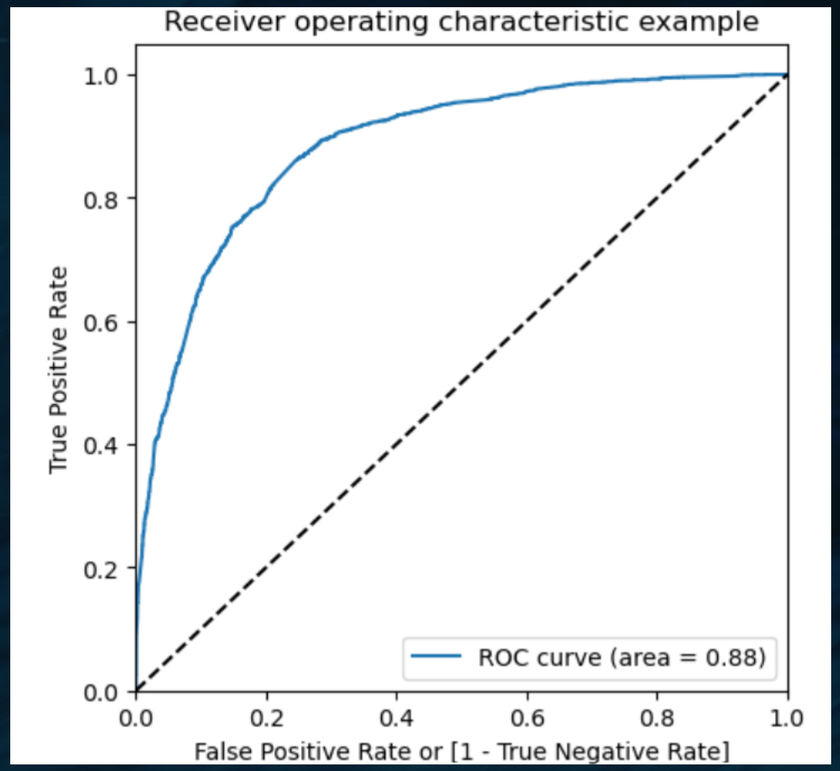
BOX PLOT



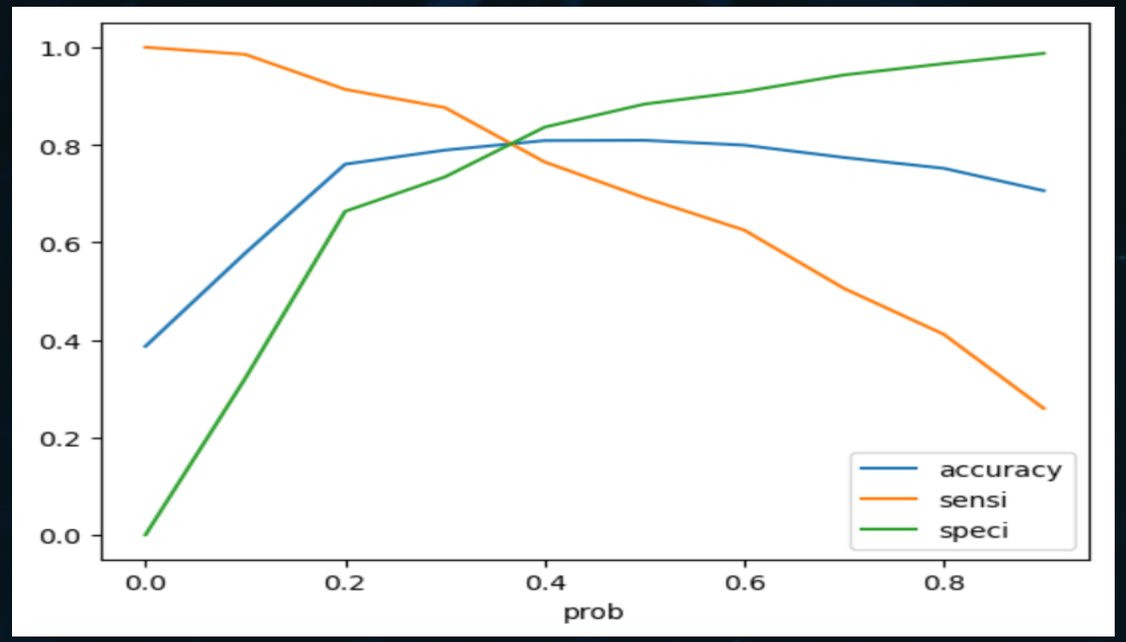
HEAT MAP

ROC Curve

1.



2.



- Finding Optimal Cut off Point
- Optimal cut-off probability is that Probability where we got balanced sensitivity and specificity.
- From the second graph its visible that optimal cut-off is at 0.35

Prediction ON TEST SET

- Before predicting on the test set, we need to standardize the test set and need to have exact same columns present in our final train dataset.
- After doing the above step, we started predicting the test set, and the new prediction values were saved in a new data frame.
- After this we did model evaluation, finding accuracy, precision, and recall.
- The accuracy score we found was 0.80, precision 0.75, and recall 0.75 approximately.
- This shows that our test prediction is having accuracy, precision, and recall scores in an acceptable range.
- This also shows that our model is stable with good accuracy and recall/sensitivity.
- Lead Score is created on test dataset to identify hot leads – high the lead score higher the chance of conversion, low the lead score lower the chance of getting converted.

Model Building:

1. Splitting the Data into Training and Testing Sets.
2. The first basic step for regression is performing a train-test split, we have chosen 70:30 Ratio.
3. Use RFE for Feature Selection.
4. Running RFE with 15 variables as output.
5. Building Model by removing the variable whose p-value is greater than 0.05 and vi value is greater than 5.
6. Prediction on test data set.
7. Overall accuracy 80%.

Conclusion

It was found that the variables that mattered the most in the potential buyers are (In descending order) :

- ✓ The total time spent on the Website.
- ✓ Total number of visits.
- ✓ When the lead source was Google,
  Direct Traffic,
  Organic Search,
- ✓ When the last activity was SMS,
  Olark chat conversation.
- ✓ When the lead origin is Lead add format.
- ✓ When their current occupation is as a working professional.

Keeping these in mind X Education can flourish as they have a very high chance to get almost all the potential buyers to change their mind and buy their course.

Thank You

Thank you for your time and consideration. We appreciate your interest in our work.

A **by Arup Ghosh**